DOCUMENT RESUME

ED 429 086 TM 029 617

AUTHOR Mittag, Kathleen C
TITLE A National Survey of AERA Members' Perceptions of the Nature and Meaning of Statistical Significance Tests.
PUB DATE 1999-04-00
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Canada, April 19-23, 1999).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Attitudes; Beliefs; National Surveys; *Researchers; *Statistical Significance; *Statistics; *Teachers; Test Use
IDENTIFIERS *American Educational Research Association

ABSTRACT

A national survey of a stratified random sample of members of the American Educational Research Association was undertaken to explore perceptions of contemporary statistical issues, and especially of statistical significance tests. The 225 actual respondents were found to be reasonably representative of the population from which the sample was drawn. The respondents had sophisticated understanding of some statistical issues (e.g., that nonsignificant results may still be important), but other features of the perceptions (e.g., perceptions of stepwise analysis) were not as encouraging. (Contains 1 table, 9 figures, and 28 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

mittsurv.wpl 3/15/99

ED 429 086

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Kathleen Mittag
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

A National Survey of AERA Members' Perceptions of the Nature and Meaning of Statistical Significance Tests

Kathleen C. Mittag

University of Texas at San Antonio

Kathleen C. Mittag
Division of Education
University of Texas at San Antonio
6900 North Loop 1604 West
San Antonio, TX 78249

Paper presented at the annual meeting of the American Educational Research Association, Montreal, April 22, 1999. I appreciate the very helpful comments of Bruce Thompson on a previous draft of this paper.

TM029617

BEST COPY AVAILABLE

Abstract

A national survey of a stratified random sample of AERA members was undertaken to explore perceptions of contemporary statistical issues, and especially of statistical significance tests. The actual respondents were found to be reasonably representative of the population from which the sample was drawn. The respondents had sophisticated understanding of some statistical issues (e.g., that nonsignificant results may still be important), but other features of the perceptions (e.g., perceptions of stepwise analysis) were not as encouraging.

Almost as soon as statistical significance tests were popularized near the turn of this century, critics emerged (Berkson, 1938; Boring, 1919). And the criticism since then has been fairly continual (e.g., Carver, 1978; Meehl, 1978; Rozeboom, 1960). But recent commentary has been particularly striking (cf. Cohen, 1994; Kirk, 1996; Schmidt, 1996; Thompson, 1996, 1999).

These criticisms of statistical tests have provoked some advocacy for the continued use of the tests, though even most advocates concur that the tests are sometimes misused or misunderstood (e.g., Cortina & Dunlap, 1997; Frick, 1996; Robinson & Levin, 1997). Indeed, several empirical studies have shown that many researchers do not fully understand the statistical tests that they employ (Nelson, Rosenthal & Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman & Rosenthal, 1993). Thus, Tryon (1998) recently lamented,

> [T]he fact that statistical experts and investigators publishing in the best journals cannot consistently interpret the results of these analyses is extremely disturbing. Seventy-two years of education have resulted in minuscule, if any, progress toward correcting this situation. It is difficult to estimate the handicap that widespread, incorrect, and intractable use of a primary data analytic method has on a scientific discipline, but the deleterious effects are doubtless

substantial... (p. 796)

The present study was undertaken to explore current perceptions of AERA members' regarding statistical significance tests. It is not clear whether AERA members believe that statistical significance tests should be banned, should be used as before, or whether the use of such tests should be supplemented with other methods. I also explored perceptions regarding score reliability and stepwise methods, about which there have also been some continuing controversies (e.g., Thompson, 1995).

Methods

Sample

I drew a stratified random sample of roughly 4% of the AERA members listed in the membership directory. The sample was stratified by AERA divisions to insure representativeness across the 12 divisions. A total of 1,127 surveys were mailed. Some of the surveys (i.e., 90) were returned for lack of a forwarding address.

Instrumentation

A short two-sided one-sheet survey was distributed. The survey asked the gender of the participants, whether they had completed a doctoral program, and what was their primary work setting. The survey then presented 29 items to which participants responded on a one-to-five Likert scale. Except for the first five of the 29 items, which were general in nature, the items were randomly ordered and some items were reverse-worded so as to

minimize response set influences.

Results

Nonresponse Bias

A total of 246 surveys were returned. However, 21 surveys (225 / 1037 = 21.7%) were unusable due to items being omitted. As Kerlinger (1986, p. 380) noted, survey mail response rates are often about 30%. The critical question when such response rates are realized is whether the respondents are still representative of the population to which the researcher wishes to generalize. The response profiles can be analyzed to provide at least some insight regarding this important issue.

As reported in Table 1, the distribution of the 225 respondents from the 51 postal locations closely matched (r=.90) the proportional distribution of AERA members across the 50 states. The distribution of the respondents across the 12 divisions in the sample also closely matched the distribution of all AERA members (r=.89). Thus, the sample appeared reasonably representative, at least as regards these characteristics.

Roughly half (49.3%) of the respondents were males. Most of the participants (83.6%) had earned a doctoral degree. The respondents' work settings were: university (65.8%), business (9.3%), school district (8.9%), and other (16.0%).

Nine Perception Clusters

The 29 items evaluated nine clusters of perceptions. The responses to the items are presented using 95% confidence intervals about the item means.

First, Figure 1 presents responses to the first five items, which measured general perceptions on statistical issues and the on-going statistical significance testing controversy. As Figure 1 illustrates, the participants were in general agreement that researchers should use the phrase "statistically significant," rather than "significant," to describe their results. This view is consistent with the position taken on this matter by Thompson (1996).

The participants also agreed that this controversy is likely to continue "for many years in the future." And they disagreed rather strongly with the proposition that statistical significance tests should be banned; at least some scholars have argued in favor of such a ban (e.g., Carver, 1978; Schmidt, 1996).

Second, participants were asked about their perceptions of the General Linear Model. Statisticians have long argued that all parametric methods are part of a single family, and that all are correlational (e.g., Cohen, 1968; Knapp, 1978). As reported in Figure 2, the respondents were basically neutral on the point of whether all analyses are correlational. However, they agreed that regression can be used to test hypotheses about means.

Third, the participants were asked whether stepwise methods identify the best variable set and whether the results can be used to infer variable importance. As reported in Figure 3, these two views were not rejected by the respondents. Statisticians tend to argue that both views should be resoundingly rejected

(Huberty, 1989; Thompson, 1995). For example, Cliff (1987, p. 185) noted that "most computer programs for [stepwise] multiple regression are positively satanic in their temptations toward Type I errors," and that "a large proportion of the published results using this method probably present conclusions that are not supported by the data" (pp. 120-121).

Fourth, the respondents were asked their views regarding score reliability. For example, as reported in Figure 4, they did not reject out of hand the view that tests are reliable *per se* in favor of a more thoughtful view that the same test may yield different reliability coefficients upon each administration (Vacha-Haase, 1998). Indeed, the respondents had neutral views of all four items dealing with score reliability.

Fifth, the participants were asked their views regarding Type I and Type II errors. For example, as reported in Figure 5, the respondents tended to disagree that "a Type II error is impossible if the results are statistically significant." Yet, one can only make a Type I error if results are statistically significant.

Sixth, the participants were asked their perceptions regarding the influence of sample sizes on statistical tests. As reported in Figure 6, for example, respondents were fairly neutral regarding whether "Significance tests are partly a test of whether the researcher had a large sample," and "Every null hypothesis will eventually be rejected at some sample size."

Seventh, the participants were asked their views of whether

statistical probabilities are exclusively measures of effect size. As reported in Figure 7, the respondents were not inclined to agree that p values directly measure study effect size or that the failure to obtain statistical significance means that results were not noteworthy or important. This is certainly a heartening finding.

Eighth, the participants were asked about the perceptions of p values as direct measures of result value. Happily, as reported in Figure 8, the participants strongly agreed that "Studies with non-significant results can still be very important."

Lastly, the respondents were asked about whether p values evaluate population parameters and result replicability. Of course, the objection that statistical significance tests do not evaluate result replicability has been central to some recent criticisms (Cohen, 1994; Sohn, 1998; Thompson, 1996). Unfortunately, as reported in Figure 9, the respondents' views on this point were fairly neutral.

Discussion

As Pedhazur and Schmelkin (1991) noted, "probably very few methodological issues have generated as much controversy" (p. 198) as have the use and interpretation of statistical significance tests. It is not clear that the controversy has raised consciousness as regards all the related issues. The present results contain both somewhat heartening and somewhat disheartening findings.

For example, it is discouraging that more researchers do not

yet realize that stepwise methods simply do not identify the best predictor set of a given size (Cliff, 1987; Huberty, 1989). Thus, Thompson (1995) provided a heuristic demonstration of a regression analysis of four variables. Predictors 1 and 2 were entered by the stepwise analysis; however, the best predictor set of size 2 involved predictor 3 and 4, for which the R^2 value was larger. Thus, the best predictor set of size 2 did not include any of the predictors identified by stepwise!

It is also somewhat disheartening that more researchers did not realize that, because the probability of obtaining sample findings that exactly match those specified by the null hypothesis is infinitely small, for a given set of sample results the null hypothesis will always be rejected at some sample size (Thompson, 1996). As statistician Roger Kirk (1996) recently emphasized,

> It is ironic that a ritualistic adherence to null hypothesis significance testing has led researchers to focus on controlling the Type I error that cannot occur because *all* null hypotheses are false. (p. 747, emphasis added)

However, it was very heartening to find, as reported in Figure 7, that the respondents were not willing to interpret statistically nonsignificant findings as being inherently unimportant. It was also encouraging, as reported in Figure 8, that at least participants did not endorse the interpretation that statistical tests evaluate whether results are likely to replicate in future research. This erroneous interpretation has been at the core of recent criticisms of common research

practices (Cohen, 1984; Thompson, 1996).

Further movement of the field as regards the use of statistical tests may require continuing elaboration of more informed editorial policies (McLean & Ernest, 1998), because many researchers tend to do what editors expect (Kirk, 1996). Sutlive and Ulrich (1998) enumerated the ideal features of such expectations:

> Based on the recent literature of other disciplines, several recommendations for evaluating and reporting research findings are made. They include calculating and reporting effect size..., placing greater emphasis on replication of results, evaluating results in a sample size context... (p. 103)

The present results provide one snapshot of the contemporary thinking of AERA members. It may be helpful to continue to monitor the evolution in thinking, as the field continues to resolve conflicting views related to the use of statistical tests.

References

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. Journal of the American Statistical Association, 33, 526-536.

Boring, E.G. (1919). Mathematical vs. scientific importance. Psychological Bulletin, 16, 335-338.

Carver, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cliff, N. (1987). Analyzing multivariate data. San Diego: Harcourt Brace Jovanovich.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Cortina, J.M., & Dunlap, W.P. (1997). Logic and purpose of significance testing. Psychological Methods, 2, 161-172.

Frick, R.W. (1996). The appropriate use of null hypothesis testing. Psychological Methods, 1, 379-390.

Huberty, C.J (1989). Problems with stepwise methods--better alternatives. In B. Thompson (Ed.), Advances in social science methodology (Vol. 1, pp. 43-70). Greenwich, CT: JAI Press.

Kerlinger, F. N. (1986). Foundations of behavioral research (3rd ed.). New York: Holt, Rinehart and Winston.

Kirk, R. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological

Bulletin, 85, 410-416.

McLean, J.E., & Ernest, J.M. (1998). The role of statistical significance testing in educational research. Research in the Schools, 5, 15-22.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Nelson, N., Rosenthal, R., & Rosnow, R.L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. American Psychologist, 41, 1299-1301.

Oakes, M. (1986). Statistical inference: A commentary for the social and behavioral sciences. New York: Wiley.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

Robinson, D., & Levin, J. (1997). Reflections on statistical and substantive significance, with a slice of replication. Educational Researcher, 26(5), 21-26.

Rosenthal, R., & Gaito, J. (1963). The interpretation of level of significance by psychological researchers. Journal of Psychology, 55, 33-38.

Rozeboom, W.W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57, 416-428.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. Psychological Methods, 1, 115-129.

Sohn, D. (1998). Statistical significance and replicability. Theory and Psychology, 8, 291-311.

Sutlive, V.H., & Ulrich, D.A. (1998). Interpreting statistical

significance and meaningfulness in adapted physical activity research. Adapted Physical Activity Quarterly, 15, 103-118.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. Educational and Psychological Measurement, 55, 525-534.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them?. Theory & Psychology, 9(2), 167-183.

Tryon, W.W. (1998). The inscrutable null hypothesis. American Psychologist, 53, 796.

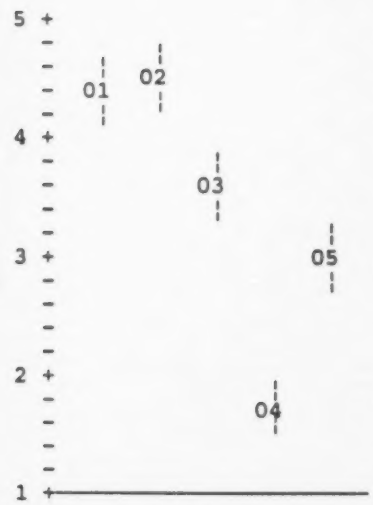
Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58, 6-20.

Zuckerman, M., Hodgins, H.S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. Psychological Science, 4, 49-53.

Table 1
Respondent Bias Evaluation Using Geography

State	Data Source: Sample	Data Source: Population
1 AK	0.45%	0.23%
2 AL	2.27%	0.98%
3 AR	0.00%	0.36%
4 AZ	0.91%	2.04%
5 CA	11.82%	11.77%
6 CO	1.82%	1.92%
7 CT	1.36%	1.60%
8 DC	1.82%	1.42%
9 DE	0.00%	0.44%
10 FL	3.64%	2.92%
11 GA	1.36%	2.75%
12 HI	0.91%	0.50%
13 IA	3.18%	1.53%
14 ID	0.00%	0.25%
15 IL	11.36%	6.43%
16 IN	2.27%	2.47%
17 KS	0.91%	0.79%
18 KY	1.82%	0.99%
19 LA	0.91%	1.35%
20 MA	1.36%	3.82%
21 MD	1.82%	2.54%
22 ME	0.91%	0.56%
23 MI	4.55%	4.22%
24 MN	1.36%	1.66%
25 MO	0.91%	1.50%
26 MS	0.45%	0.39%
27 MT	0.00%	0.16%
28 NC	4.09%	2.37%
29 ND	0.45%	0.17%
30 NE	0.45%	0.60%
31 NH	0.00%	0.42%
32 NJ	2.27%	3.01%
33 NM	0.00%	0.85%
34 NV	0.45%	0.36%
35 NY	6.82%	8.91%
36 OH	2.73%	3.89%
37 OK	1.36%	0.82%
38 OR	2.73%	1.14%
39 PN	4.55%	4.55%
40 RI	0.45%	0.49%
41 SC	0.91%	0.78%
42 SD	0.00%	0.17%
43 TN	1.36%	1.63%
44 TX	7.27%	4.97%
45 UT	0.45%	0.84%
46 VA	3.18%	2.92%
47 VT	0.00%	0.33%
48 WA	1.36%	2.06%
49 WV	0.00%	0.44%
50 WI	0.91%	2.50%
51 WY	0.00%	0.19%

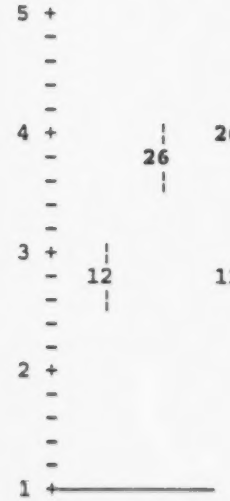
Figure 1
General Views Regarding Statistical Testing



2. It would be better if everyone used the phrase, "statistically significant," rather than "significant", to describe the results when the null hypothesis is rejected.
1. Controversies regarding the use of significance tests have existed for many years in the past, and will doubtless continue for many years in the future.
3. Most studies are conducted with insufficient statistical power against Type II error.
5. All that significance means is that the researcher rejected the null hypothesis.
4. Science would progress more rapidly if tests of significance were banned from journal articles.

Note. 5 = "Agree"; 1 = "Disagree".

Figure 2
Perception of the General Linear Model

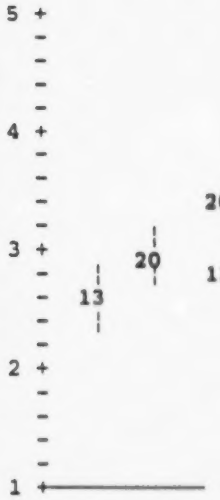


26. *It is not possible to use regression to statistically test the null that means of different groups are equal.*

12. All statistical analyses (e.g., t-tests, ANOVA, r, R) are correlational.

Note. For item 26, 1 = "Agree," and 5 = "Disagree." For item 12, 1 = "Disagree," and 5 = "Agree."

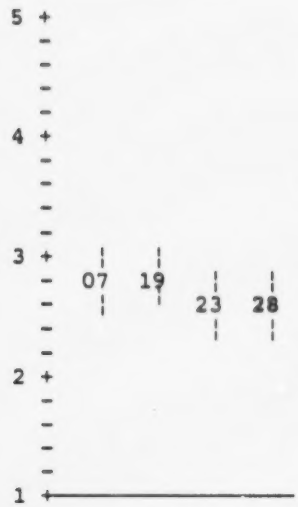
Figure 3
Perceptions of Stepwise Analysis



20. *When researchers do stepwise analyses, the order of the entry of the variables (1st, 2nd, etc.) provides one useful indication of the importance of the variables.*
13. *In regression and other analyses, stepwise analyses can reasonably be used to identify the best subset of predictors of a given subset size.*

Note. For both items, 1 = "Agree," and 5 = "Disagree."

Figure 4
Perceptions of Score Reliability

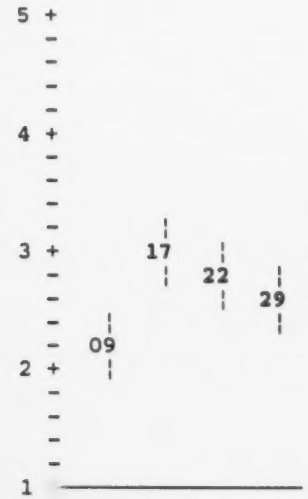


19. Testing the significance of a reliability or a validity coefficient with a null hypothesis that $r^2=0$ is not useful or productive.
7. On its face, the statement, "the reliability of the test," asserts an untruth, since reliability is not a characteristic of a given test.
23. Poor reliability of data in a given study will tend to lower or attenuate the effect sizes that are detected.
28. *Reliability does <u>not</u> directly affect the likelihood of obtaining significance in a given study.*

Note. For items 7, 19, and 23, 1 = "Disagree," and 5 = "Agree." For item 28, 1 = "Agree," and 5 = "Disagree."

BEST COPY AVAILABLE

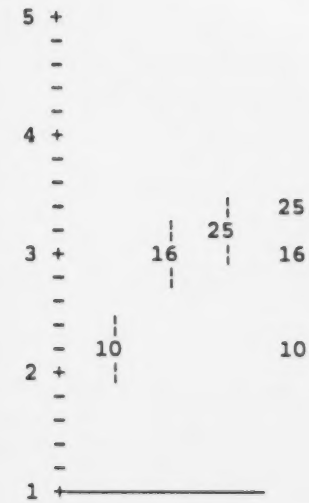
Figure 5
Perception of Type I and Type II Errors



17. *Type I errors may be a concern when the null hypothesis is <u>not</u> rejected.*
22. *It is possible to make both a Type I and a Type II error in a given study.*
29. *Type II errors are probably fairly common within published research.*
9. A Type II error is impossible if the results are statistically significant.

<u>Note</u>. For items 17, 22 and 29, 1 = "Agree," and 5 = "Disagree." For item 9, 1 = "Disagree," and 5 = "Agree."

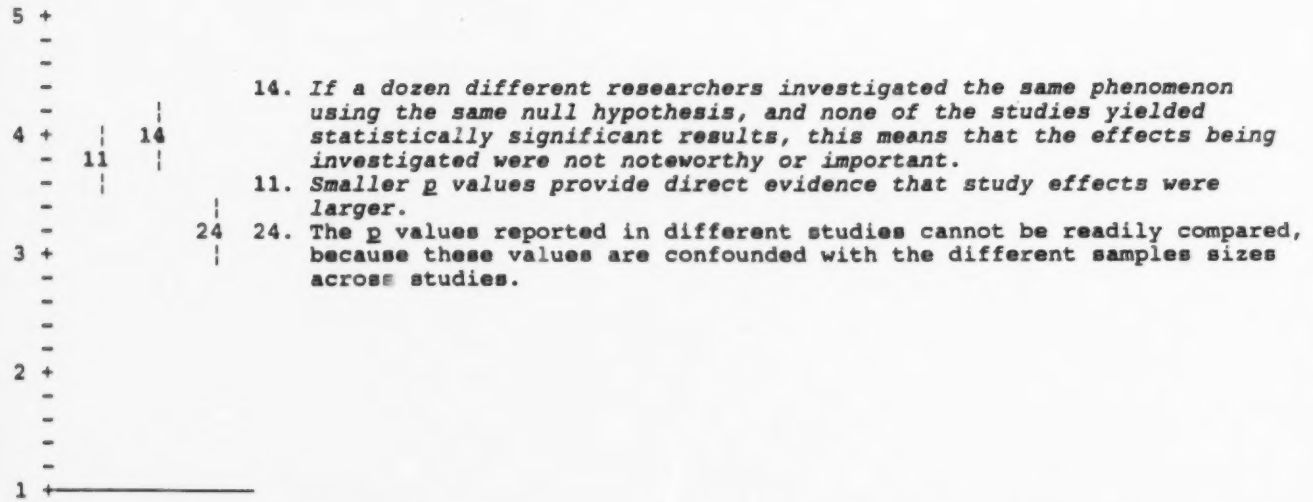
Figure 6
Perceptions of Sample Size Influences



25. Significance tests are partly a test of whether the researcher had a large sample.
16. Every null hypothesis will eventually be rejected at some sample size.
10. Statistically significant results are more noteworthy when sample sizes are small.

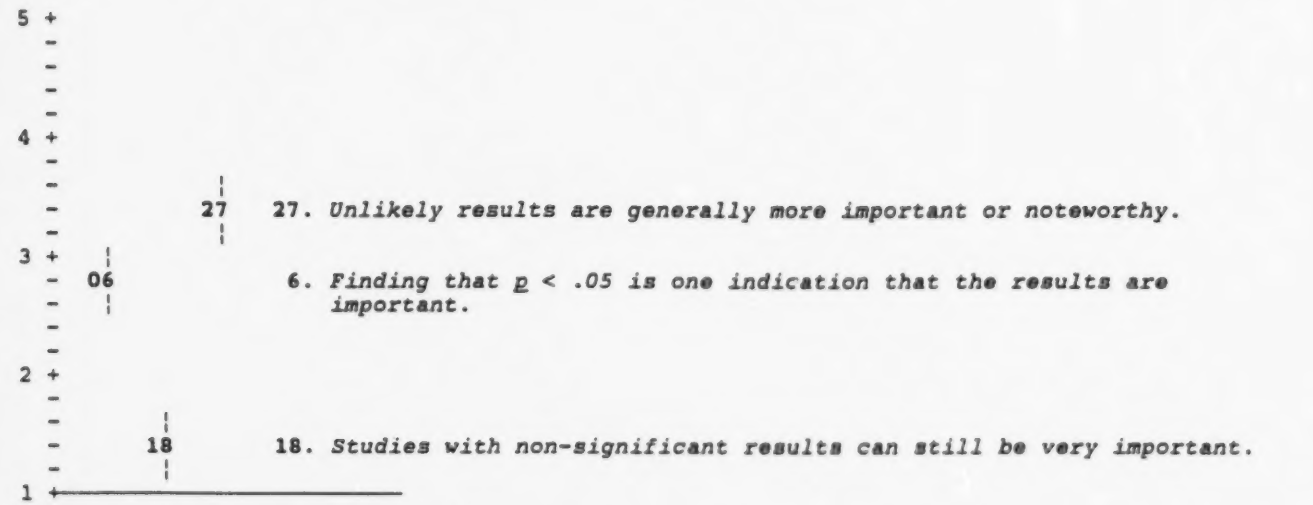
<u>Note</u>. 5 = "Agree"; 1 = "Disagree".

Figure 7
Perceptions of Effects


5
4
3
2
1
14
11
24
14. If a dozen different researchers investigated the same phenomenon using the same null hypothesis, and none of the studies yielded statistically significant results, this means that the effects being investigated were not noteworthy or important.
11. Smaller p values provide direct evidence that study effects were larger.
24. The p values reported in different studies cannot be readily compared, because these values are confounded with the different samples sizes across studies.

Note. For items 11 and 14, 1 = "Agree," and 5 = "Disagree." For item 24, 1 = "Disagree," and 5 = "Agree."

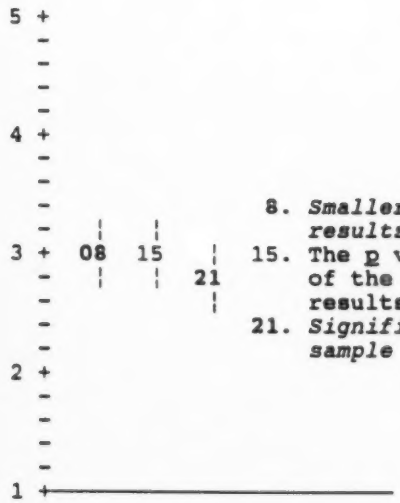
Figure 8
Perceptions of p as Importance



Note. 1 = "Agree"; 5 = "Disagree".

BEST COPY AVAILABLE

Figure 9
Perceptions of p as Replicability Evidence



8. *Smaller and smaller values for the calculated p indicate that the results are more and more likely to be replicated in future research.*
15. The p values that are calculated in a given study test the probability of the results occurring in the sample, and not the probability of results occurring in the population.
21. *Significance tests evaluate the probability that the results for the sample are the same in the population.*

Note. For items 8 and 21, 1 = "Agree," and 5 = "Disagree."

BEST COPY AVAILABLE